SPECIFIC AIMS

During the pilot phase of the Illuminating the Druggable Genome (IDG) program, our team published over 43 peer-reviewed papers that cite the grant, and developed new online software systems that have already been widely accessed. Our leading achievement during the pilot phase of IDG was the development, deployment, and publication of the Harmonizome resource [1]. The Harmonizome is a database with a web-based user interface providing access to ~72 million associations between genes/proteins and functional biological and biomedical terms, collected from 66 publicly available resources that were converted into 114 datasets. The Harmonizome resource provides landing pages for genes, gene-sets (biological terms), datasets, and resources. Since September 2015, more than 160,000 unique users have visited the Harmonizome website, based on Google Analytics. The Harmonizome resource epitomizes the IDG concept because its gene/protein landing pages illuminate knowledge about all human genes and proteins, including those under-studied potential targets listed in the RFA [2]. The Harmonizome resource organizes and abstracts data from many omics- and literature-based resources that profile all human genes and proteins, and this enables us not only to serve this knowledge, but also to impute knowledge using machine learning (ML). New knowledge about the functions of under-studied kinases, ion channels, and GPCRs can be predicted by ML; for example, knockout mice phenotypes or association with diseases could be predicted. Building on our successes from the IDG pilot phase, during the implementation phase of IDG, we plan to continue our efforts in the same direction. We will continue to process and abstract knowledge about genes/proteins, gene-sets/pathways, drugs/small-molecules, diseases/phenotypes, and cells/tissues from over 100 resources and databases. We will expand efforts to impute knowledge with ML. In addition, we will work closely together, and with the Data and Resource Generation Centers (DRGCs), the other KMC team, and the Resource Dissemination and Outreach Center (RDOC). The overall work of our team will synergistically enhance our knowledge about human genes and proteins, and the online software and database resources that we will produce will become instrumental to the work of many biomedical researchers. To achieve these ambitious goals, we will follow these specific aims:

Aim 1. Abstract, process, analyze, and integrate data from over 100 publicly available primary resources that produce new information about genes and proteins with a focus on illuminating knowledge about the understudied kinases, GPCRs, and ion channels

A. Systematically continue the abstraction and integration of data from all possible resources by processing the data into attribute tables, gene-sets, bi-partite graphs, and functional association networks. The data processing scripts will be shared as Jupyter notebooks [3] and on GitHub [4]. This data will produce knowledge about associations between genes/proteins and the biological and biomedical functional terms that involve them. These efforts will enrich the content of the gene set enrichment analysis for Enrichr [5], MSigDB [6], and GSEA-P [7].

B. Once all this data is organized into these data structures, apply supervised and unsupervised machine learning methods to unravel gene/protein, gene-set/pathway, drug/small-molecule, cell/tissue, and patient/disease/side-effect networks; and impute functions for the under-studied targets, including associations with drugs, pathways, cells/tissues, and diseases. Such learning will be interactive, dynamic, reproducible, and will enable complex user queries.